The Annals of Applied Statistics
2016, Vol. 10, No. 1, 365–394
DOI: 10.1214/15-AOAS897

PRECINCT OR PREJUDICE? UNDERSTANDING RACIAL DISPARITIES IN NEW YORK CITY'S STOP-AND-FRISK POLICY

BY SHARAD GOEL, JUSTIN M. RAO AND RAVI SHROFF

Stanford University, Microsoft Research and New York University

Recent studies have examined racial disparities in stop-and-frisk, a widely employed but controversial policing tactic. The statistical evidence, however, has been limited and contradictory. We investigate by analyzing three million stops in New York City over five years, focusing on cases where officers suspected the stopped individual of criminal possession of a weapon (CPW). For each CPW stop, we estimate the ex ante probability that the detained suspect has a weapon. We find that in more than 40% of cases, the likelihood of finding a weapon (typically a knife) was less than 1%, raising concerns that the legal requirement of "reasonable suspicion" was often not met. We further find that blacks and Hispanics were disproportionately stopped in these low hit rate contexts, a phenomenon that we trace to two factors: (1) lower thresholds for stopping individuals—regardless of race—in high-crime, predominately minority areas, particularly public housing; and (2) lower thresholds for stopping minorities relative to similarly situated whites. Finally, we demonstrate that by conducting only the 6% of stops that are statistically most likely to result in weapons seizure, one can both recover the majority of weapons and mitigate racial disparities in who is stopped. We show that this statistically informed stopping strategy can be approximated by simple, easily implemented heuristics with little loss in efficiency.

1. Introduction. Over the last 10 years, New York City residents have been stopped and briefly detained by the police millions of times in an effort to get weapons, drugs and other contraband off the streets. Proponents of this stop-question-frisk policy (hereafter called "stop-and-frisk") argue that by strictly enforcing weapon and drug possession laws, one indirectly reduces more serious crime, such as murder and armed robbery, in line with the "broken windows" theory of policing [Wilson and Kelling (1982)]. Though it is difficult to rigorously assess this claim, wide adoption of stop-and-frisk by the New York City Police Department (NYPD) in the early 1990s did coincide with a period of substantial decline in crime in the city. Opponents of stop-and-frisk, however, argue that regardless of whether the policy is effective, it violates two constitutional protections. First, they claim individuals are stopped without legal basis, in violation of the Fourth Amendment. Indeed, in nearly 90% of cases, stopped suspects are

Received January 2015; revised September 2015.
Key words and phrases. Criminology, discrimination, racial profiling, risk assessment, Fourth Amendment.

released without any further action, suggesting that the vast majority of individuals stopped were not engaged in serious criminal activity.[1] Second, they claim the policy is not applied in a race-neutral manner, in violation of the Fourteenth Amendment. Notably, blacks and Hispanics make up more than 80% of individuals stopped, even though they constitute approximately 50% of the New York City population.

By and large, nearly all academic research on stop-and-frisk and related tactics has focused on claims of racial discrimination. However, there is little consensus in the literature on the magnitude—or even the existence—of such discrimination, even among papers that study the same policy in the same city over the same time frame. Moreover, almost no statistical attention has been paid to possible Fourth Amendment violations. In an effort to cast further light on this ongoing statistical and policy debate, we analyzed three million stops conducted by New York City police officers between 2008 and 2012, one of the largest studies of stop-and-frisk to date. Of these three million stops, we focus our attention on the approximately 760,000 instances in which an individual was detained under suspicion of criminal possession of a weapon (CPW), in part because the success of these stops is readily determined by the presence or absence of a weapon.

We make three main contributions. First, we develop a novel statistical and legal approach to detecting and assessing possible Fourth Amendment violations in stop-and-frisk. The Fourth Amendment requirement of "reasonable suspicion" for police stops was established in Terry v. Ohio (1968), and subsequently expanded on in several court rulings, including Illinois v. Wardlow (2000).[2] As established in these rulings, reasonable suspicion exists when there are *articulable* facts or circumstances which would lead a reasonable person to suspect that a crime has been, is being or will be committed—a standard of proof lower than probable cause but higher than a mere hunch. Random searches, regardless of whether they are an effective deterrent, are generally prohibited under the Fourth Amendment.[3] To determine whether this threshold has been met, we estimate the ex ante likelihood (i.e., the likelihood based only on information available to officers prior to the stop decision) that the stopped individual has a weapon. We find that in 43% of the approximately 300,000 CPW stops between 2011 and 2012, there was at most a 1% chance of finding a weapon on the suspect.[4] We note that the recovered

[1]As we discuss in detail below, we would not expect all legally conducted stops to result in an arrest, citation or other such disciplinary action. Moreover, some stopped individuals may in fact have been found to be engaged in criminal activity (e.g., trespass), but officers were able to resolve the situation without further formal police action.

[2]Stop-and-frisk has a complicated legal history, which we only briefly address in this paper. For a more comprehensive review, see Gelman, Fagan and Kiss (2007).

[3]There are some exceptions to this general rule. For example, the U.S. Supreme Court has found sobriety checkpoints to be constitutional [Michigan Dept. of State Police v. Sitz (1990)].

[4]Stops from 2008–2010 are used to train our statistical models, and thus not included in this tally.

weapons are typically knives, with guns constituting approximately 10% of found weapons. Whether these stops in fact violate the Fourth Amendment is a complex legal question, and one that is largely outside the scope of this paper. Nevertheless, our results do suggest that individuals were often stopped with relatively little evidence of criminal activity, corroborating recent court rulings rebuking the NYPD for its stop-and-frisk tactics [Davis v. City of New York (2013), Floyd v. City of New York (2013), Ligon v. City of New York (2013)].

Second, we find that blacks and Hispanics were disproportionately involved in low hit rate stops, and building on work by Gelman, Fagan and Kiss (2007), we trace this disparity to two factors: (1) the highly localized nature of the policy, and (2) discriminatory enforcement. Specifically, we find that high crime areas, particularly public housing, have lower stop thresholds, presumably reflecting more aggressive efforts to reduce crime in those locations. Since these areas are home to large numbers of blacks and Hispanics, members of these groups were disproportionately impacted by stop standards that differed by location. After correcting for these highly localized policing tactics—as well as adjusting for several other factors—we find that stopped blacks and Hispanics were less likely than similarly situated whites to possess a weapon, suggestive of racial discrimination in stop decisions. We note that without understanding the location-specific nature of stop-and-frisk, it is easy to conflate racial discrimination with generally low—but not necessarily discriminatory—stop thresholds in predominately minority neighborhoods.

Finally, we show that by conducting only the highest ex ante hit rate stops, one can dramatically reduce the overall number of stops while largely preserving the number of successful ones. In particular, we show that one can recover 50% of weapons by conducting only the 6% of CPW stops with the highest ex ante hit rate, and 90% of weapons by conducting 58% of CPW stops. These ex ante hit rates are based only on information observable to officers prior to the stop decision, and so it is at least in theory possible to implement such a strategy. Further, since low hit rate stops disproportionately involve blacks and Hispanics, optimizing for weapons recovery would simultaneously bring more racial balance to stop-and-frisk. To facilitate adoption of such strategies by police departments, we develop stop heuristics that approximate our full statistical model via a simple scoring rule. Specifically, we show that with a rule consisting of only three weighted stop criteria, one can recover the majority of weapons by conducting 8% of stops.

Related work. Given the significance and salience of stop-and-frisk, a number of statistical studies have assessed various aspects of the issue, particularly claims of racial bias, which we briefly review. In an early, comprehensive analysis, Gelman, Fagan and Kiss (2007) concluded that minorities were stopped more often than whites, both in comparison to their proportion in the local population and relative to local crime rates in those groups. A subsequent analysis also found evidence of racial disparities, but concluded that the magnitude of the effect was

relatively small, and in particular estimated that only 15 out of 3000 NYPD officers stopped an unusually high number of black and Hispanic suspects [Ridgeway (2007), Ridgeway and MacDonald (2009)]. Coviello and Persico (2013) fit an economic model of behavior to the stop-and-frisk data and found no evidence of racial bias. Finally, investigating the ramifications of local events on policing, Legewie (2016) showed that in the days following fatal shootings of two NYPD officers by black suspects, there was an increase in the use of physical force against blacks—but not whites or Hispanics—during stops; moreover, such increase in force was not observed after the murder of two police officers by a white and a Hispanic suspect.

Several authors have also studied the closely related issue of racial discrimination in traffic stops. In a novel design, Grogger and Ridgeway (2006) analyzed traffic stops in Oakland and showed that the racial distribution of stopped individuals during the day, when a suspect's race is readily apparent, matches the distribution at night, when the "veil of darkness" masks race, and thus concluded there was little bias in stop decisions. However, Ridgeway (2006) finds differences in post-stop outcomes by race; for example, black drivers are less likely than whites to have stops lasting less than ten minutes. Knowles, Persico and Todd (2001) distinguish between so-called statistical and taste-based discrimination [Arrow (1973)], and do not find evidence of racial prejudice against blacks in Maryland traffic stops. Examining traffic stops by the Boston Police Department, however, Antonovics and Knight (2009) show that officers are more likely to conduct a search if the race of the officer differs from the race of the driver, consistent with taste-based racial discrimination. Anwar and Fang (2006) show that such tests based on officer race can be misleading; they introduce an alternative statistical method and do not find evidence of discrimination in stops carried out by the Florida Highway Patrol. Finally, Epp, Maynard-Moody and Haider-Markel (2014) trace the extensive and complex history of race and police stops in the United States.

2. Data and methods.

2.1. *Data description*. Our primary dataset consists of all 2.9 million stops conducted and recorded by the New York City Police Department (NYPD) between January 1, 2008, and December 31, 2012. Following a stop, officers complete a UF-250 stop-and-frisk form, recording various aspects of the stop, including demographic characteristics of the suspect, the time and location of the stop, the suspected crime and the rationale for the stop (e.g., whether the suspect was wearing clothing common in the commission of a crime). One notable limitation of this dataset is that no demographic or other identifying information is available about officers.

After an individual is stopped, officers may conduct a frisk (i.e., a quick pat-down of the person's outer clothing) if they reasonably suspect the individual is armed and dangerous; officers may additionally conduct a search if they have

TABLE 1
Summary of key information recorded on the UF-250 stop-and-frisk form

Field	Value
Date	yyyy-mm-dd
Time	hh:mm
Location	GPS coordinates
Precinct	1–123
Location type	Public housing, public transit or neither
Inside or outside	Inside or outside
Suspect's sex	Male or female
Suspect's race	White, black, Hispanic, Asian or other
Suspect's build	Heavy, medium, muscular or thin
Suspect's age	Integer (years)
Suspect's height	Integer (inches)
Suspect's weight	Integer (pounds)
Observation period	Integer (minutes)
Officer in uniform	Yes or no
Radio run	Yes or no
Suspected crime	1 of 113 prespecified categories (e.g., criminal possession of a weapon and robbery)
Primary stop circumstance(s)	Suspicious object, fits description, casing, acting as lookout, suspicious clothing, drug transaction, furtive movements, actions of violent crime, suspicious bulge and/or other
Additional stop circumstance(s)	Witness report, ongoing investigation, proximity to crime scene, evasive response, associating with criminals, changed direction, high crime area, time of day, sights and sounds of criminal activity and/or other
Suspect frisked	Yes or no
Suspected searched	Yes or no
Suspect arrested	Yes or no
Weapon found on suspect	Yes or no
Drugs found on suspect	Yes or no

probable cause of criminal activity. Frisks and searches occur in 56% and 9% of cases, respectively. An officer may decide to make an arrest (6% of instances) or issue a summons (6% of instances), all of which is recorded on the UF-250 form. Responses are subsequently standardized, compiled and released annually to the public. A list of key information collected is summarized in Table 1.

While officers are mandated to complete a UF-250 form for investigations initiated based on reasonable suspicion, they may not always do so, and so a possibly large number of stops go undocumented. Also, there is evidence that officers follow "scripts of suspicion" when filling out forms to justify stops [Fagan and Geller (2014)]. Further, it is not always even clear whether a police encounter formally constitutes a "stop." (The legal test is whether a reasonable person would not have

felt free to terminate the encounter, though there is at times genuine ambiguity with this criterion.) Finally, since these forms are completed by hand, there are likely errors in recording and transcribing stop details. Our dataset is thus neither a complete nor fully accurate record of all conducted stops. Nevertheless, we note that in light of recent litigation [Daniels et al. v. the City of New York (2001)], the NYPD now works to ensure UF-250 accuracy, including supervisor review. Moreover, these data (and related datasets) have been used in past academic work [Gelman, Fagan and Kiss (2007), Ridgeway (2007), Ridgeway and MacDonald (2009)] and in a variety of high-profile court cases, including Floyd v. City of New York (2013). As such, we assume the data are generally suitable for our analysis, and we note the effect of possible problems with the data on our results where appropriate.

A common metric for evaluating stop-and-frisk is the so-called hit rate, the proportion of stops in which a suspect was arrested, a summons issued or some other outcome occurred that suggests the guilt of a stopped individual. Hit rates are regularly used to assess the level of proof applied when stopping a suspect, with lower hit rates corresponding to less stringent standards [Ayres (2002), Becker (1993, 2010)]. In particular, lower arrest rates for stopped blacks relative to stopped whites are often interpreted as indicating that the threshold for stopping blacks is lower than for stopping whites, consistent with claims of racial discrimination. However, as noted in Gelman, Fagan and Kiss (2007), one could reasonably reach the opposite conclusion: relatively higher arrest rates of whites could indicate that officers are biased *against* whites in that they arrest them too often.

To circumvent these issues of interpretation, we first subset the data to include only the 760,502 stops between 2008 and 2012 for which the suspected crime was listed as criminal possession of a weapon (CPW), by far the most commonly occurring suspected crime in our dataset. We note that officers are required to articulate the suspected crime prior to conducting the stop, though this information, along with all other stop details, is recorded afterward. Then, instead of considering whether or not the stopped individual was arrested, we look at whether the suspect was found to have a weapon, which is also recorded on the UF-250 form. This approach has three advantages. First, relative to arresting a suspect, there is arguably less officer discretion involved in determining whether an individual has a weapon.[5] Second, the presence or absence of a weapon directly indicates whether the stop was ex-post justified under the explicitly stated suspicion of CPW. In contrast, a stopped individual could be arrested for a variety of reasons (e.g., drug possession) that are unrelated to the original purpose of the stop. Finally, by focusing on a single class of well-defined suspected crimes, we mitigate ecological fallacies due to different base hit rates for various crime categories. For example, since hit rates are generally higher when the suspected crime is drug related, and since a relatively higher proportion of whites are involved in these drug stops, one could reach spurious conclusions by aggregating all stops.

[5]We indeed see that not all suspects found to have a weapon are arrested.

2.2. *Estimating stop-level hit rates.* As discussed above, estimating race-specific hit rates helps both to assess whether stops meet the standard of reasonable suspicion and also to test for racial discrimination. Traditionally, hit rates are estimated by simply computing the overall percentage of stops among each race group that result in the outcome of interest (e.g., finding a weapon on or arresting the stopped suspect), possibly controlling for the distinct contexts (e.g., time of day) in which individuals of different races are stopped [Gelman, Fagan and Kiss (2007), Ridgeway (2007)]. In contrast to these aggregate, race-level hit rate statistics, our aim is to estimate stop-level hit rates. Specifically, our primary quantity of interest is the ex ante likelihood that any given CPW stop results in finding a weapon on the stopped suspect. That is, at the moment an officer decides to stop an individual for suspicion of criminal possession of a weapon, we seek the probability—taking all information available to the officer at the time—that the suspect has a weapon. This methodological approach has two advantages. First, by computing the full hit rate distribution, we can estimate the fraction of CPW stops that fall below a given evidence threshold (e.g., where the likelihood of finding a weapon is less than 1%), which in turn helps to assess possible violations of Fourth Amendment protections against unreasonable search. Second, stop-level probabilities can be efficiently aggregated to estimate hit rates for various small subgroups (e.g., stopped Hispanics in a given precinct), circumventing issues of data sparsity and allowing us to quantify the extent to which the threshold for stopping individuals differs across contexts.

To compute stop-level hit rates, we first fit a logistic regression model on the 301,513 CPW stops between 2009 and 2010 with complete UF-250 forms, where the left-hand side is the probability of finding a weapon on the stopped suspect and the right-hand side includes several variables recorded on the form that would have been available immediately before the stop. Specifically, we include indicator variables for the suspect's demographics (sex, race and build); whether the stop occurred on public transit, in public housing or neither; whether the stop occurred inside or outside; the date and time of the stop (month, day of week and time of day, binned into disjoint four-hour blocks); one or more reasons for the stop (e.g., furtive movements and high crime area, as detailed in Table 1); whether the stop was the result of a radio run; and whether the officer was in uniform. We additionally include continuous variables for the year, suspect's height, weight and age, and the time for which the officer observed the suspect before stopping him or her (the latter four are all normalized to have mean 0 and variance 1).

We further include in the model two location-specific features: (1) indicator variables for the precinct where the stop occurred; and (2) local hit rate, as described below. Together, these geographic features help account for both local crime rates and enforcement standards that differ by location. For each year t and location s, the local hit $h_t(s)$ is the weighted percentage of CPW stops during year t that result in the recovery of a weapon, where stops are weighted according to

their distance from s. Specifically,

$$h_t(s) = \sum_{i=1}^{n_t} y_i \exp\left(-\frac{d(s,s_i)^2}{2}\right) \Big/ \sum_{i=1}^{n_t} \exp\left(-\frac{d(s,s_i)^2}{2}\right), \tag{1}$$

where n_t is the total number of CPW stops during year t, $y_i \in \{0,1\}$ indicates whether the ith stop was successful (i.e., whether a weapon was recovered), s_i is the location of the ith stop, and $d(s,s_i)$ is the geodesic distance in kilometers between s and s_i. Such Gaussian kernel averaging is a standard approach for estimating local intensities in spatial processes [Diggle (1985)]. In our model, each stop occurring in year t at location s is annotated with the feature $h_{t-1}(s)$. Accordingly, when estimating the ex ante likelihood of stop success, we only assume the previous year's statistics are available, avoiding look-ahead bias.

Finally, we include in the model all pairwise interactions between these variables (including self-interactions). Thus, the final form of the model is

$$\mathbb{P}(y_i = 1) = \operatorname{logit}^{-1}\left(\sum_k \alpha_k x_{k,i} + \sum_{k \le \ell} \beta_{k,\ell} x_{k,i} x_{\ell,i}\right), \tag{2}$$

where y_i indicates whether the ith stop resulted in finding a weapon on the suspect, $x_{k,i}$ denotes features of the stop, and α and β are the model coefficients. The model is trained on the 301,513 CPW stops from 2009–2010, with data from 2008 additionally used to generate the necessary local hit rate statistics. Out-of-sample predictions are then produced for the 288,158 CPW stops from 2011–2012, and it is this set of stops that we primarily use in our subsequent analysis.

Given the large number of stops (301,513) and variables (7705, including interactions) that we consider, we fit the logistic regression model (2) with stochastic gradient descent (SGD).[6] Stochastic gradient descent is a highly scalable method popular in the machine learning community for its speed and low use of memory. In contrast to traditional gradient descent, SGD streams through the data and, on each iteration, computes an approximate gradient estimated from the current datapoint. Implicit regularization is obtained by stopping the optimization procedure after a single pass through the data, before full convergence has occurred [cf. Bottou (1998)].

To provide some insight into which features the model makes the most use of, we list in Table 2 the positive and negative coefficients with largest absolute value.

[6]We use the open-source package Vowpal Wabbit (VW) with the default values for all algorithm parameters; in particular, we fit the model with a single streaming pass through the data. Because the fitted model depends on the order of the examples, we separately train the model on 100 random shufflings of the data and average the results. We explored several other model fitting techniques, including L^2-regularized logistic regression as implemented in Scikit-learn [Pedregosa et al. (2011)], and found VW performed best in terms of both speed and model fit. Notably, VW consistently produced more calibrated predictions than the alternatives.

TABLE 2
The ten positive and negative model coefficients with largest absolute value

Coefficient	Value
(Local hit rate) × (precinct 73)	0.53
(Local hit rate) × (precinct 33)	0.53
(Location = neither) × (suspicious object)	0.44
(Location = transit) × (precinct 73)	0.43
(Location = housing) × (suspicious object)	0.43
(Local hit rate) × (precinct 60)	0.40
(Location = transit) × (radio run)	0.39
(Local hit rate) × (precinct 52)	0.39
(Suspicious object) × (suspect sex = male)	0.38
Suspicious object	0.36
(Precinct 69) × (suspicious clothing)	−1.30
(Precinct 114) × (suspected drug transaction)	−1.24
(Precinct 49) × (Monday)	−1.22
(Precinct 114) × (acting as lookout)	−1.16
(Precinct 71) × (suspicious clothing)	−1.03
(Precinct 114) × (August)	−1.03
(Precinct 101) × (Thursday)	−1.02
(Precinct 109) × (suspected drug transaction)	−1.00
(Precinct 70) × (suspected drug transaction)	−0.99
(Precinct 42) × (suspect race = other)	−0.99

In particular, many of the highest weighted features are location-specific, a point we return to below. While we believe it is helpful to inspect the features in order to gain intuition about the model, we stress that the features should not be interpreted in terms of their statistical significance, and we therefore intentionally do not include standard errors. In strongly regularized models, such interpretations can be misleading since bias in coefficient estimates is often larger than the variance indicated by the standard errors [Kyung et al. (2010)].

The strength of our conclusions rests on the accuracy of our model, and so we examine model performance in several ways. First, on the test set of 2011–2012 CPW stops, we find AUC is 83%, indicating high out-of-sample performance. We next check calibration by comparing the model-predicted probabilities to the empirical hit rates. Figure 8(a) in the Appendix confirms the model is well calibrated along the entire range of predicted probabilities. We further compare the model estimates to the empirical hit rates for various subgroups of the population—including categories defined by age, race, gender and location—and likewise find the model performs well [Figures 8(c)–(d) in the Appendix]. Finally, we repeat our primary analysis with a random forest classifier, considered to be one of the best statistical methods for large-scale classification [Fernández-Delgado et al. (2014)].

In the Appendix, we show that random forest yields estimates largely in line with those from logistic regression.[7] Last, we note that although a suspect's height, weight and age can only be approximated by the officer before the stop, the fitted model is largely robust to reasonable errors in these terms. In particular, if we assume officers estimate height, weight and age with independent, mean zero errors that are normally distributed with standard deviations of 2 inches, 10 pounds and 5 years, respectively, the mean absolute change in estimated ex ante probability is 0.1 percentage points. The totality of evidence thus suggests our modeling framework produces accurate and robust estimates.

3. Results.

3.1. *Assessing reasonable suspicion.* With the fitted model described by equation (2) in hand, we assign each CPW stop from 2011–2012 a model-inferred ex ante probability of finding a weapon on the stopped suspect. Figure 1(a) shows the distribution of these ex ante probabilities for the approximately 290,000 CPW stops in this time period. As indicated by the dotted vertical line, the overall likelihood of finding a weapon is 3%. Moreover, 43% of the stops had less than a

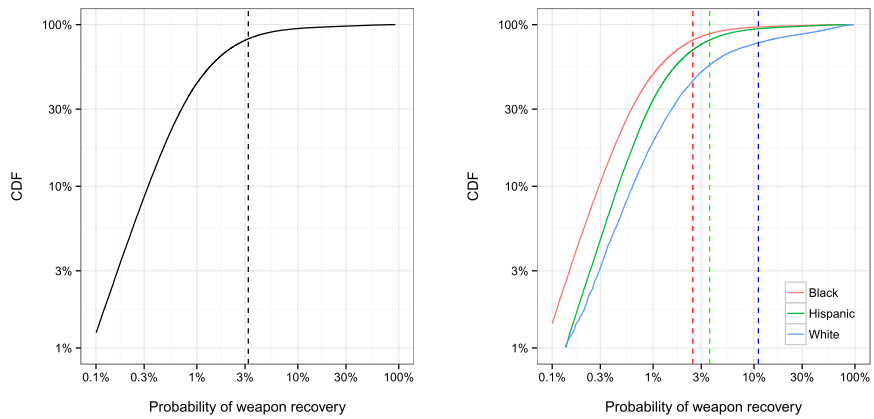

FIG. 1. *Distribution of the ex ante probability of finding a weapon on a suspect stopped for suspicion of criminal possession of a weapon (CPW). Panel* (a) *shows the distribution over all such stops between* 2011 *and* 2012, *with the vertical line indicating the overall likelihood of finding a weapon on a stopped suspect.* 43% *of all CPW stops have less than a* 1% *ex ante chance of turning up a weapon. Panel* (b) *disaggregates this distribution by suspect race, where the vertical lines show the likelihood of finding a weapon on black, Hispanic and white suspects. Stopped blacks and Hispanics are much less likely to have a weapon than stopped whites.*

[7]We ultimately used logistic regression for our primary analysis since it was considerably faster and gave somewhat more calibrated estimates.

1% chance of turning up a weapon, and 19% of the stops had less than a 0.5% chance. Though the courts have yet to quantify the standard of "reasonable suspicion" for stop-and-frisk in terms of precise probabilistic thresholds [Rudovsky and Rosenthal (2013)], our results indicate that a substantial fraction of CPW stops are conducted on the basis of relatively little evidence.[8]

As a point of comparison, in the landmark New York City stop-and-frisk court case, Floyd v. City of New York (2013), reasonable suspicion was assessed by hand-classifying each possible stated justification for the stop (as indicated on the UF-250 form) as reasonable or not. For example, whereas "furtive movements" in the absence of any other indicator of criminality was deemed insufficient justification, "furtive movements" together with "high crime area" was deemed acceptable. Though that analysis resulted in 5% of all stops (including non-CPW stops) classified as unreasonable, the presiding judge in the case believed the classification overly conservative, and suggested the true number of stops lacking reasonable suspicion was likely considerably higher [Floyd v. City of New York (2013), page 41]. In contrast, while our purely statistical approach admittedly does not explicitly consider legal precedent, it does offer a straightforward, fast and largely objective method for directly relating reasonable suspicion to criminality.

Figure 1(b) shows stop-level hit rate distributions broken down by the race of the stopped suspect (black, Hispanic or white), with the vertical lines indicating overall hit rates for each race group. In particular, consistent with past results [Gelman, Fagan and Kiss (2007)], the overall hit rates for blacks and Hispanics (2.5% and 3.6%, resp.) are considerably lower than for whites (11%). In other words, these results indicate that when blacks and Hispanics are stopped, it is typically on the basis of less evidence than when white suspects are stopped. Moreover, while 49% of blacks stopped under suspicion of CPW have less than a 1% chance of in fact possessing a weapon, the corresponding fraction for Hispanics is 34%, and is just 19% for stopped whites. Thus, if we equate reasonable suspicion with a particular probability threshold (say 1%), a far greater fraction of stops of blacks and Hispanics are unwarranted than are stops of whites.

3.2. *Heterogeneity in hit rate by location.* It is perhaps tempting to conclude that the lower hit rates of blacks (2.5%) and Hispanics (3.6%) relative to whites (11%) is indicative of racial discrimination. However, as Ridgeway (2007) points out, whites and minorities are typically stopped in different contexts, and so differing hit rates may not be the result of racial bias. Indeed, as we discuss below, stop-

[8]One could interpret "reasonable" as meaning a hit rate higher than the base rate in the general population, and since New York City has some of the most restrictive gun laws in the country, a 1% hit rate may in fact be an order of magnitude or more higher than that. However, such a threshold seems overly tolerant, as the standard set out in *Terry* requires one to reasonably suspect a person has been, is or is about to be engaged in criminal activity. Nevertheless, the ultimate standard of reasonable suspicion is for the courts and legal scholars to determine.

and-frisk is an extremely localized tactic, heavily concentrated in high-crime, predominantly black and Hispanic areas, and so lower tolerance for suspicious activity (and hence lower hit rates) in these areas could account for the racial disparity.

Figure 2(a) shows the distribution of CPW stops in 2011–2012 colored by the race of the stopped suspect (a random sample of 10,000 stops is plotted), illus-

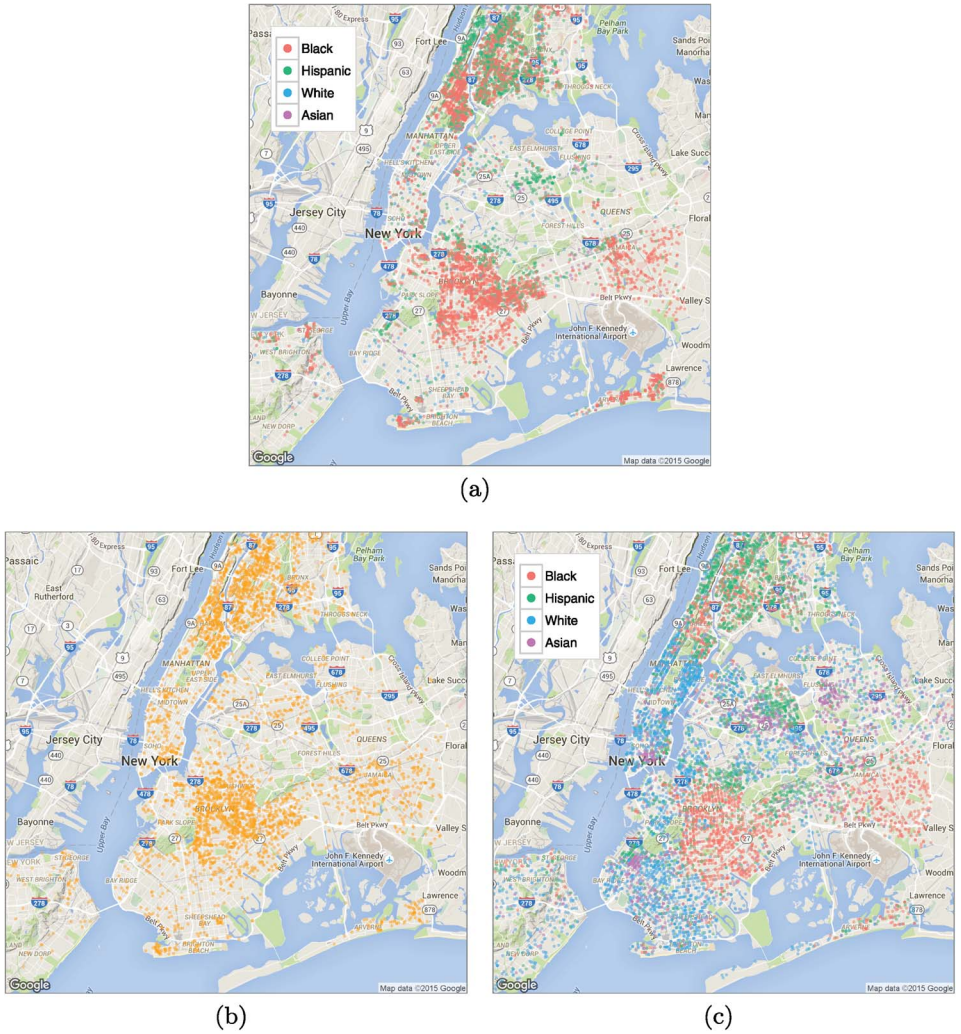

FIG. 2. *Panel* (a) *shows the geographic distribution of CPW stops between* 2011 *and* 2012, *colored by the suspect's race. For comparison*, *Panel* (b) *shows the distribution of murders in New York City between* 2006 *and* 2011, *which indicates that stop-and-frisk is primarily employed in high-crime areas. Finally*, *Panel* (c) *shows the racial distribution of the general population based on* 2010 *block--level U.S. Census data*, *highlighting that these high-crime*, *high stop-and-frisk areas are disproportionately black and Hispanic.*

trating how geographically specific the use of stop-and-frisk is. For comparison, Figure 2(b) plots each recorded homicide in New York City from 2006–2011, 2427 in total, as compiled by the New York Times.[9] The distribution of homicides is remarkably well aligned with the distribution of stops, indicating that the NYPD concentrated its use of stop-and-frisk on high-crime areas. Finally, Figure 2(c) shows the distribution of the general New York City population, by race, based on 2010 block-level U.S. Census data. To generate the plot, 10,000 individuals were sampled from Census records and placed on the map at the middle of their Census block, the smallest geographic unit for which information is publicly available.

The maps in Figure 2 highlight three points. First, there is an almost one-to-one correspondence between areas with heavy use of stop-and-frisk [Figure 2(a)] and areas with high incidence of violent crime [Figure 2(b)]. While this is a natural and possibly effective policing strategy, a consequence of the tactic is that individuals who live in high-crime areas, but who are not themselves engaged in criminal activity, bear the costs associated with being stopped. Second, these high-crime areas are overwhelmingly black and Hispanic. Accordingly, the cost of stop-and-frisk is largely shouldered by minorities. Third, by comparing Figure 2(a) and (c), we see that the racial composition of stopped individuals is similar to the racial composition of the neighborhoods in which stop-and-frisk is heavily employed. Thus, the striking racial composition of stopped CPW suspects (61% are black, 30% are Hispanic and 4% are white) appears at least qualitatively attributable to selective use of stop-and-frisk in minority-heavy areas, illustrating the importance of understanding the localized nature of the policy.

Adding quantitative detail to these qualitative results, we estimate hit rates for each of the 77 precincts in New York City and further distinguish between stops occurring in public housing, on public transit or in other locations (primarily pedestrian stops). Figure 3(a) shows the results, plotting the hit rate of white versus black suspects for each location, with the size of the points indicating the number of stops.[10] To generate the estimates, the ex ante probabilities from equation (2) are averaged over stops in each geographic area; for areas with a large number of stops, the model agrees with the simple, empirical hit rate, but the model-estimated statistics lead to more stable estimates for the sparser regions.

As indicated by the plot, there is substantial variation in average hit rate across locations, ranging from less than 1% in some public housing units to more than 30% for transit stops in certain precincts. Moreover, within region, though the hit rates of white and black suspects are not identical, they are much more similar than the city-wide averages (indicated by the dashed horizontal lines). Specifically,

[9] See http://projects.nytimes.com/crime/homicides/map for further details. Only data up until 2011 were available.

[10] We only plot those precinct/location-type combinations with at least 10 black and 10 white stops, accounting for 96% of all stops.

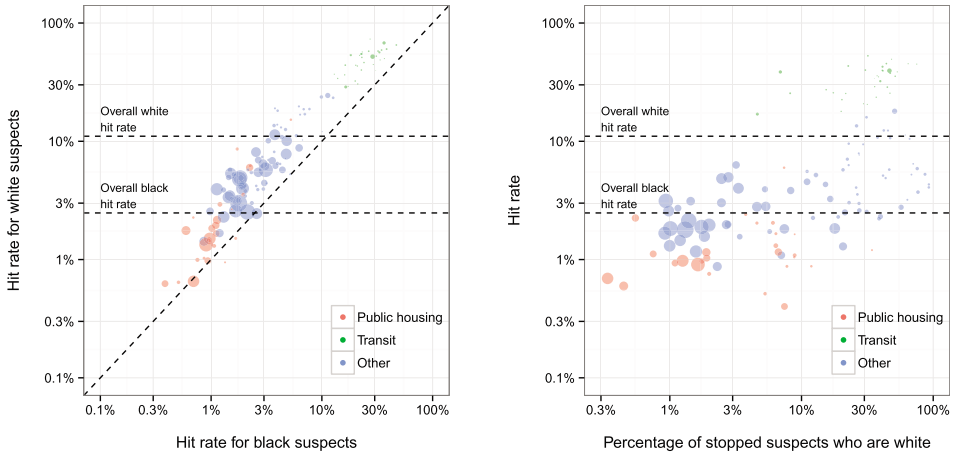

FIG. 3. *Likelihood of finding a weapon* (*hit rate*) *on suspects stopped for suspicion of CPW by geographic area, where we consider only stops of suspects who are either black or white. Each circle corresponds to stops within a given precinct in one of three possible location types, indicated by their shading*: *public housing, public transit or other* (*typically street stops*). *The areas of the circles indicate the number of stops in that location. Panel* (a) *compares the hit rate among black and white suspects for each area, showing that the within-area hit rates for the two groups are much more similar than the overall group averages. Panel* (b) *plots each area's hit rate by its percentage of stopped suspects who are white* (*among white and black suspects*), *and indicates that low hit rate stops generally occur in areas where primarily black suspects are stopped, and, moreover, such areas account for a large fraction of total stops.*

the average within-area ratio of white hit rate to black hit rate, weighted by the number of stops in each area, is 2.0.[11] By comparison, the city-wide ratio is 4.5. Thus, a significant fraction of the racial disparity in hit rates can be explained by policing tactics that vary considerably by area.

Figure 3(b) further illustrates this point, plotting for each area its overall hit rate (irrespective of race) by the percentage of stopped suspects who are white (among suspects who are either white or black). We again see that predominately black areas are associated with low hit rates, while predominately white areas have high hit rates, further demonstrating the importance of location for understanding the adverse effects of stop-and-frisk on minorities.

While much of the racial disparity in hit rates is explained by geography, we note that this finding is orthogonal to the question of whether stops meet the stan-

[11]This weighted average is $\sum_{i=1} w_i r_i / \sum_i w_i$, where r_i = white hit rate/black hit rate in area i, w_i is the number of stops in area i, and i ranges over all precinct/location-type combinations with at least one black and one white stop.

dard of reasonable suspicion, as discussed in Section 3.1.[12] Indeed, it appears that stops in several public housing complexes have quite low average hit rates (less than 1%), calling into question that the bar for reasonable suspicion has been met. Corroborating our statistical findings, recent stop-and-frisk lawsuits have revealed that NYPD training materials explicitly instructed officers to question people in New York City Housing Authority buildings "without reasonable suspicion of trespass, and to arrest for trespass those who fail to leave or affirmatively establish their right" to be present in a building [Davis v. City of New York (2013)].

3.3. *Testing for racial discrimination.* Although much of the racial disparity in hit rates disappears once we account for the location of a stop, hit rates for whites are still consistently higher than for blacks across geographic area, leaving open the possibility that racial bias is still at play. Location, however, is not the only possible confounding factor. For example, the demographic composition of the local populations could shift with the time of day, aligning with patrol schedules to affect race-specific hit rates. Alternatively, the distribution of age may vary across race, with certain age groups—and consequently race groups—more often the target of low hit rate stops.[13]

To adjust for these alternative, nonrace-based explanations, we use the logistic regression model described above in equation (2) to estimate the hit rate of hypothetical *similarly situated* whites. Namely, for each of the 178,742 CPW stops of blacks between 2011 and 2012 with no missing information, we use the model to generate the ex ante likelihood of finding a weapon on the stopped suspect assuming the suspect was white, but preserving all other aspects of the stop. We note that because the model includes a number of interaction terms, this estimate does not simply differ from the original by a constant factor, but rather depends on the precise combination of features describing the stop.

The result of this exercise is displayed in Figure 4, where we plot the hit rate for stopped blacks against the hit rate for similarly situated whites, grouped by location. The plot shows that by adjusting for the various differences in context between stops of whites and blacks, we do indeed shrink the hit rate gap. We also find, however, that the gap does not disappear, with the overall hit rate of similarly situated whites about 50% larger than the black hit rate (3.8% compared to 2.5%), indicated by the dashed lines. Further, the higher white hit rate is not simply due

[12]One may be tempted to conclude that having stop thresholds that vary by location in and of itself indicates the reasonable suspicion standard is violated. Why, one might argue, should one precinct's bar for reasonable suspicion differ from an adjacent precinct's? The law, however, requires only that a minimum standard of proof be met, and a precinct may choose not to stop all individuals above that legal threshold for a variety of legitimate reasons, including constrained resources and alternative police priorities.

[13]Such stop policies could still be illegal if stop thresholds were based on a protected class, such as age.

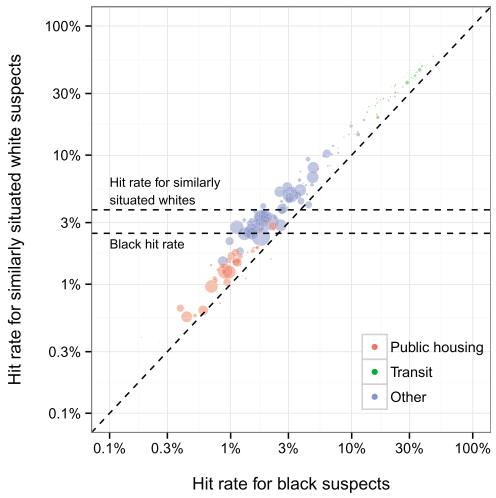

FIG. 4. *The likelihood of finding a weapon* (*hit rate*) *among black suspects stopped for suspicion of CPW compared to model-estimates for similarly situated white suspects*, *disaggregated by geographic area. Each circle corresponds to stops within a given precinct in one of three possible location types*, *indicated by their shading*: *public housing*, *transit or other* (*typically street stops*). *Even after adjusting for a variety of features of the stop—including time of day*, *physical characteristics of the suspect and officers' stated justification for conducting the stop—the hit rate among whites* (3.8%) *is still higher than among blacks* (2.5%).

to a few anomalous areas, but holds consistently across nearly every location we consider. It thus appears that relative to similarly situated whites, black suspects are indeed stopped with less ex ante evidence of a crime, corroborating claims of racial discrimination.

Such racial discrimination could in principle arise from two qualitatively distinct mechanisms—statistical or taste-based [Arrow (1973), Ewens, Tomlin and Wang (2014), Persico (2009)]—which our analysis cannot disentangle. With statistical discrimination, officers may genuinely believe that blacks are more likely to carry weapons than the data suggest, perhaps due to faulty heuristics or limited opportunity to estimate event probabilities. For example, an object that is considered "suspicious" on a black individual may not be considered "suspicious" on a white person [Eberhardt et al. (2004)]. In contrast, with taste-based discrimination, officers may accurately estimate ex ante hit rates, but apply a lower standard of proof when stopping blacks than whites.

While we see that racial disparities in stop-and-frisk are in part driven by discrimination, variation in local stop thresholds still appears to be a primary driver of disparate racial impact. For example, consider the 42,941 blacks who were stopped in housing projects in 2011–2012, a subgroup with overall hit rate of 1%. Now, if we suppose those individuals were white—but otherwise identical—we estimate a hypothetical hit rate of 1.3%, where the increase is indicative of

racial discrimination. If, however, we suppose those 42,941 black individuals were stopped on the street, as opposed to in housing, with all other traits—including race and precinct—kept identical, we estimate a hypothetical hit rate of 2%, higher than both the actual hit rate (1%) and the hit rate (1.3%) of similarly situated whites.

3.4. *Improving stop efficiency.* As we have seen, individuals are regularly stopped under suspicion of CPW in contexts where it is unlikely that they in fact possess a weapon. This observation begs the question, how can we design a better policy? Statistical risk assessment has a long history in criminal justice [Berk (2012)]. For example, by analyzing 1.5 million pretrial records, the Arnold Foundation recently developed a model to estimate the likelihood that a defendant released before trial will engage in violence, commit a new crime or fail to return to court [Milgram et al. (2015)]. More than 20 cities and states have adopted this tool to help judges decide which defendants to detain and which to release.

Building on this tradition, we use the statistical model in equation (2), trained on the 301,513 CPW stops in 2009–2010, to estimate the ex ante probability that a stopped suspect has a weapon. This procedure yields the following family of stop rules. For any threshold $p > 0$, stop an individual if: (1) the individual would have been stopped under the usual stop-and-frisk practice; and (2) the probability of recovering a weapon, as estimated under the model, is at least p. The first condition is critical since the model is trained only on stops that in fact occurred, and so it may not generalize to the population at large. One can thus think of this as a two-step procedure, where an officer first relies on his or her usual training to determine whom to possibly stop, and then checks whether the model-estimated probability exceeds a prespecified threshold, set perhaps by the city or police department.[14]

To evaluate the performance of this approach, we first use the model to estimate the ex ante likelihood that each of the 288,158 CPW stops in 2011–2012 would turn up a weapon. We then rank stops in descending order by this likelihood, with the stops deemed most likely to result in finding a weapon accordingly appearing at the top of the list. We note that this ranking is based on out-of-sample predictions and, moreover, only uses data available at the moment right before an officer decides to stop an individual. Finally, since, for these stops, we know whether or not a weapon was ultimately found on the suspect, we can estimate how many weapons one would have recovered had only the top x-percent of stops been conducted.

[14]By appropriately setting the stop threshold, one can balance the asymmetric costs of false positives and false negative. For example, setting a high threshold would lower the number of false positives while raising the number of false negatives. It may be possible to obtain improved performance by explicitly specifying a loss function that takes into account the asymmetric costs, and then directly optimizing the stop decision for this loss [Bach, Heckerman and Horvitz (2006), Berk (2012)].

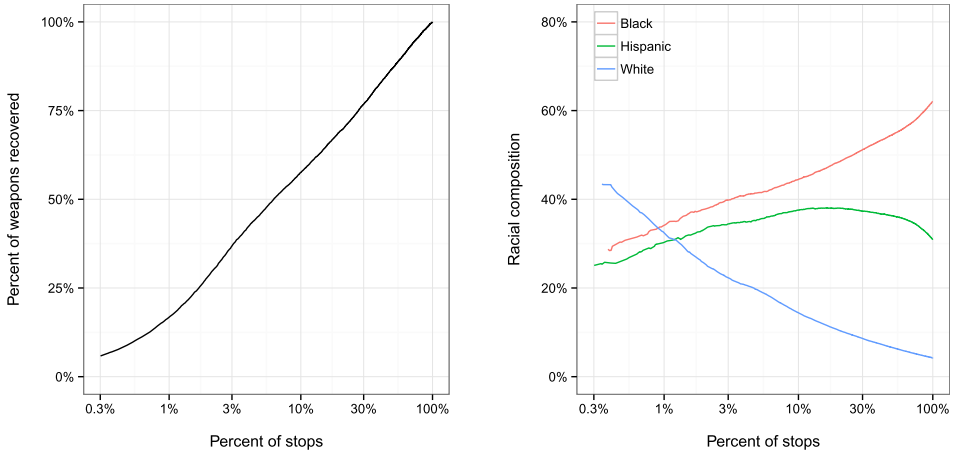

FIG. 5. *Panel* (a) *plots the estimated percentage of weapons recovered as a function of the number of stops conducted, where stops are ordered by their model-predicted likelihood of turning up a weapon, from highest to lowest.* (*Note that the* x*-axis is on a log scale.*) *In particular, the best* 10% *of stops result in* 58% *of weapons recovered, and the best* 50% *result in* 87% *of weapons recovered. Panel* (b) *shows how the racial composition of stopped suspects varies with the number of stops, where stops are again ordered from most to least likely to result in turning up a weapon. Since low likelihood stops disproportionately involve black suspects, reducing the number of stops results in lowering the overall proportion of stopped suspects who are black.*

Figure 5(a) shows this curve, where we normalize the number of recovered weapons on the y-axis by the total number of weapons recovered in all CPW stops from 2011–2012. Remarkably, we find that only 6% of stops are needed to recover the majority of weapons, and only 58% are necessary to turn up 90% of the weapons. Because so few CPW stops have any significant chance of turning up a weapon, we can eliminate a large number of stops and still identify almost as many individuals who are carrying weapons. Since a disproportionate number of the lowest hit rate stops involve blacks and Hispanics, eliminating these stops also alters the racial composition of stopped suspects, as shown in Figure 5(b). For example, whereas blacks make up 61% of all CPW stops in 2011–2012, they comprise 44% of the 10% of stops most likely to result in finding a weapon. A more racially-balanced pool of stopped suspects could temper public reaction to the policy, including resentment and distrust of the police [Lerman and Weaver (2014)]. One can thus view this to be an added benefit of improving stop efficiency.

Since late 2013, the use of stop-and-frisk in New York City has been severely curtailed, both because of several court rulings critical of stop-and-frisk as well as a newly elected mayor, Bill de Blasio, who openly opposes the tactic. Specifically, during the last four months of 2013, there were 3985 CPW stops, compared to 33,683 for the same period in 2012, a reduction of 88%. Are officers system-

atically conducting only the "best" stops (i.e., those stops most likely to result in finding a weapon on the suspect)? We find the CPW hit rate for the end of 2013 is substantially higher, 11%, as compared to the same period in 2012, 3%. However, had the officers conducted the 3985 stops ranked highest by our model, we would expect a hit rate of 17%. It thus seems that while the NYPD is indeed focusing on higher hit rate stops, there is still considerable room for improvement by rigorously optimizing the policy.

3.5. *Heuristic stop strategies*. The strategy of conducting only the exante most efficient stops is conceptually simple, but it is admittedly not straightforward to implement in practice. Officers cannot simply evaluate a complex statistical model in their heads when deciding whether or not to stop a suspect (although technology, such as a handheld computer, could help with this). Further, it seems unlikely that police departments would adopt an opaque machine learning model to inform stop decisions. To address these difficulties, we draw on a large body of work which has found that simple, transparent and interpretable heuristics often work as well as complex statistical models [Czerlinski, Gigerenzer and Goldstein (1999), Gigerenzer and Goldstein (1996), Lovie and Lovie (1986), Ustun and Rudin (2014)].

To start, as in equation (2), we model the likelihood of recovering a weapon in a CPW stop via logistic regression. This time, however, we use only the 18 stop circumstances officers already consider (listed in Table 1, excluding the two "other" categories), indicator variables for each of the 77 precincts and indicator variables for the three location types (public housing, transit and "neither"); we do not include interactions. To further reduce model complexity and increase interpretability, we constrain the 18 coefficients corresponding to stop reasons to be non-negative. This non-negativity constraint captures the intuitively reasonably assumption that all else equal, the 18 stop factors only increase the likelihood an individual has a weapon. For example, regardless of the stop location or which other stop circumstances are recorded, those with a "suspicious bulge" should presumably be more likely to have a weapon than those without.[15] We thus fit the *reduced model*

$$\mathbb{P}(y_i = 1) = \text{logit}^{-1}\left(\sum_{j=1}^{18} \alpha_j a_{j,i} + \sum_{k=1}^{77} \beta_k b_{k,i} + \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i}\right) \tag{3}$$

with the constraint $\alpha_j \geq 0$, where a, b and c are indicator variables for stop reason, precinct and location type, respectively.[16]

[15] In an unconstrained model, we find that several of the 18 reasons are in fact negative. However, the unconstrained model performs only marginally better than the constrained version, and so, prioritizing model simplicity and interpretability, we opt for the latter.

[16] The logistic regression coefficients were computed with the *penalized* package in R [Goeman (2010)], which provides maximum likelihood estimates for the model coefficients constrained to

TABLE 3

Values for the five nonzero coefficients for stop circumstances in the reduced model and the corresponding score for the heuristic model. In deciding whether to make a stop, officers add the relevant heuristic scores and check whether the sum exceeds an area-specific threshold

Coefficient	Value	Heuristic score
Suspicious object	2.6	3
Sights and sounds of criminal activity	0.8	1
Suspicious bulge	0.6	1
Ongoing investigation	0.1	
Witness report	0.1	

Only 5 of the 18 stop circumstances were found to have positive weight (the remaining were identically zero): (1) suspicious object; (2) sights and sounds of criminal activity; (3) suspicious bulge; (4) witness report; and (5) ongoing investigation. Notably, all five circumstances are directly tied to criminal activity, and the more subjective conditions (e.g., "furtive movements") drop out of the model. Coefficients for these five features are listed in Table 3; as before, we do not list standard errors given that penalized methods can produce strongly biased estimates.

The reduced model in equation (3) is more transparent and interpretable than the complete statistical model in equation (2), but it is still cumbersome to evaluate on the fly. We simplify the expression in two steps. First, to implement the stopping procedure described above, we need not compute the actual probability of recovering a weapon, but can instead compute a stop score that is monotonically related to the probability. We consequently ignore the logistic transformation and simply check whether the sum of the relevant coefficients exceeds a given threshold. Second, we round the five coefficients for the stop circumstances to the nearest integer (listed in Table 3); we leave the precinct and location-type coefficients unaltered.[17] This procedure results in only three nonzero coefficients for the stop reasons: suspicious object (value = 3), sights and sounds of criminal activity (value = 1) and suspicious bulge (value = 1). Letting $\tilde{\alpha}_j$ denote the rounded coefficients, and reindexing $\tilde{\alpha}_j$ so that the first three values correspond to the nonzero values, the score S_i for the ith stop is

$$S_i = \sum_{j=1}^{3} \tilde{\alpha}_j a_{j,i} + \sum_{k=1}^{77} \beta_k b_{k,i} + \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i}. \tag{4}$$

have the specified signs. The package does not provide standard errors, as these can be misleading in strongly regularized models [Kyung et al. (2010)].

[17] We tried several different rescaling and rounding schemes and obtained similar results.

Now, suppose we have selected a stop threshold T, then the stop condition $S_i \geq T$ is equivalent to

$$\sum_{j=1}^{3} \tilde{\alpha}_j a_{j,i} \geq T_r \qquad \text{where } T_r = T - \sum_{k=1}^{77} \beta_k b_{k,i} - \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i}. \tag{5}$$

The leftmost sum is a function of the three stop reasons, and T_r is an area-specific threshold that depends only on the precinct and location type. Thus, to quickly and rigorously assess the likelihood a potential stop will lead to the recovery of a weapon, officers simply need to add at most three small, positive integers (see Table 3), and check whether the sum exceeds a fixed threshold T_r for the area they are patrolling. Since officers commonly patrol only a single area during a shift, this procedure is particularly straightforward to carry out in practice.[18]

To implement this scheme, one still needs to select a stop threshold T, which in turn determines area-specific thresholds T_r. The higher the threshold, the fewer people stopped, but also the fewer weapons recovered. Figure 6(a) plots this trade-off. For various thresholds T, we compare the percent of individuals stopped under

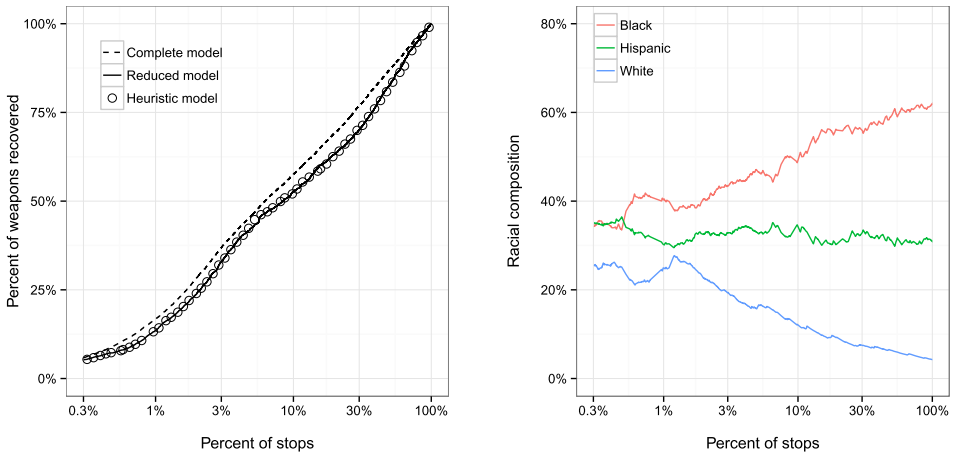

FIG. 6. *Panel* (a) *shows the estimated percentage of weapons recovered as a function of the number of stops conducted* (*note that the x-axis is on a log scale*) *for various models*. *Panel* (b) *shows how the racial composition of stopped suspects varies with the number of stops*, *where stops are ordered from most to least likely to result in turning up a weapon according to the heuristic score in equation* (4). *As in Figure* 5(b), *reducing the number of stops results in lowering the overall proportion of stopped suspects who are black*.

[18]There are a number of other heuristics one could try, including more complex policies where the stop factors vary by precinct. In the Appendix, we examine a specific alternative in which there is a uniform citywide threshold, but find it does not perform nearly as well as the one we consider here.

the heuristic model to the percent of weapons recovered, indicated by the open circles. For comparison, we also plot the trade-off under the complete model given by equation (2) and the reduced model given by equation (3). The figure shows that performance of the heuristic model is virtually indistinguishable from the reduced model. Moreover, while the heuristic model is not quite as effective as the complete model, it still performs surprisingly well. For example, with the heuristic model, one recovers 50% of weapons by making just 8% of stops; in comparison, 6% of stops are required under the complete model. Figure 6(b) shows that using the heuristic model to make fewer stops also yields a more racially balanced composition of stopped suspects. Although blacks make up 61% of CPW stops in 2011–2012, they comprise just 49% of the 10% of stops with highest heuristic score. We note that if such stop rules were ultimately adopted, the model would likely require periodic updating since changes in officers' behavior could affect model performance.

We conclude by examining the area-specific thresholds T_r. Figure 7 shows the area thresholds for a policy that recovers 50% of weapons, where higher thresholds are indicated by lighter colors.[19] A comparison with Figure 2 reveals that

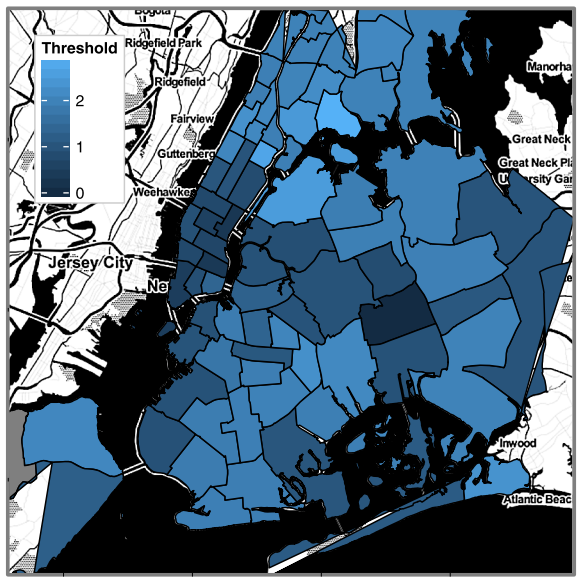

FIG. 7. *Precinct thresholds, with lighter shading indicating higher thresholds. Comparison with Figure 2 shows that precincts with high crime and high numbers of stops also tend to have high thresholds.*

[19]For ease of visualization, for each precinct we plot the average threshold over the different location types in the precinct, where the terms in the average are weighted by the number of stops in each location type.

high-crime, predominantly minority areas have relatively high thresholds. That is, in these areas, the policy requires a *higher* number of indicators of criminal activity to justify a stop. This counterintuitive observation stems from the disproportionately large number of low efficiency CPW stops in high-crime, predominantly minority neighborhoods. It could be the case that indicators of criminal behavior (e.g., "suspicious object") may not be as predictive in these neighborhoods or, alternatively, that officers patrolling such areas simply have a lower threshold for considering an object "suspicious" [Eberhardt et al. (2004)]. Finally, we note that many areas, particularly in lower Manhattan, have stop thresholds of zero. According to our stop rule, officers in such areas would thus stop suspects per their usual procedures, without additionally checking whether the stop score exceeded a threshold.

4. Discussion. By estimating the ex ante efficiency of stops, we were able to investigate claims that stop-and-frisk violated two constitutional protections: first, that individuals were detained without legal basis, in violation of the Fourth Amendment; and second, that the tactic was not applied in a race-neutral manner, in violation of the Fourteenth Amendment. Regarding the former claim, we find that in a substantial fraction of instances where a suspect was stopped for suspicion of carrying a weapon, it was in fact ex ante very unlikely a weapon would be found on the individual. In particular, in 43% of such cases, the likelihood of finding a weapon was less than 1%. Though it is beyond the scope of this paper to determine what constitutes reasonable suspicion in this context, our result raises concerns that the legal standard is often not met. Regarding the latter claim, we show that while the adverse effects of stop-and-frisk on blacks and Hispanics are largely attributable to heavy use of the tactic in high-crime, predominately minority areas, there still appears to be an element of racial bias. It is unclear whether this bias derives from racial prejudice or spurious statistical reasoning by officers. However, regardless of the underlying cause, blacks and Hispanics are subject to stops conducted on the basis of less suspicion than similarly situated whites. Finally, we show that by reducing the number of low hit rate stops—which disproportionately affect minorities due to both highly localized tactics and racial bias—one can still recover most weapons while bringing more racial balance to stop-and-frisk.

In our primary analysis, we considered only instances in which an individual was stopped for suspicion of criminal possession of a weapon, the single most frequently recorded suspected crime, constituting one-fourth of stops. Our results, though, are not just restricted to weapons possession. In particular, for stops where the suspected crime is drug related—including criminal possession and sale of marijuana and other controlled substances, comprising 10% of all stops—12% of stops in 2011–2012 have less than a 1% ex ante likelihood of contraband being found on the detained individual, and 56% have less than a 5% chance. We

likewise find that the disparate effect of such drug-related stops on blacks and Hispanics is due to a combination of highly local policing strategies and racial discrimination.

A possible objection to our approach is that even for CPW stops, recovering weapons is not the only—or perhaps not even the primary—goal of the police. Officers, for example, may simply consider stops a way to advertise their presence in the neighborhood or a means to collect intelligence on criminal activity in the area, regardless of how many weapons are directly recovered. Stops conducted for these alternative motives could quite plausibly deter individuals from carrying weapons and might lead to information helpful in solving cases, both of which presumably would lower the incidence of violent crime over time. In the instances we consider, however, the explicitly stated reason for a stop is suspicion of criminal possession of a weapon, not one of the various other reasons that may or may not withstand legal or public scrutiny, and so it seems most natural to consider whether individuals were in fact likely to be carrying weapons. Moreover, as we have previously noted, simply because a strategy may be effective does not make it legal. For instance, searching a suspect's home before a warrant is issued may be an effective way to collect evidence, but is nonetheless illegal except under exigent circumstances. A related worry is that "criminal possession of a weapon" is a catchall category for a variety of criminal offenses, and so by focusing on whether a weapon was found, we underestimate the value of a stop. Addressing this issue, we observe that our results are qualitatively similar if we instead use arrests as the outcome variable, mitigating cause for concern. Finally, one might worry that our predictive models omit key variables officers use when deciding whether to stop individuals. If so, we might overestimate the number of stops with low ex ante hit rate, and accordingly overestimate the number of potential Fourth Amendment violations. We have taken care to minimize this possibility by including hundreds of covariates that detail the circumstances of the stop. We further note that the UF-250 form that officers complete for each stop has been carefully designed to elicit and record what the NYPD believes are the important stop details.[20]

Looking forward, our results show that though stop-and-frisk does suffer from serious problems, the tactic is not beyond repair. By focusing on the relatively small number of high hit rate situations—situations that can be reliably identified via statistical analysis—one may be able to retain many of the benefits of stop-and-frisk for crime prevention while mitigating constitutional violations. This observation has the potential to not only improve New York City's stop-and-frisk program, but could also aid similar policies throughout the country.

[20] As stops require reasonable and *articulable* suspicion, departments and officers have legal incentive to record the rationale for the stop. It is a dubious legal argument to claim stops are conducted on the basis of information that cannot be articulated and recorded on the form.

APPENDIX A: MODEL CHECKS

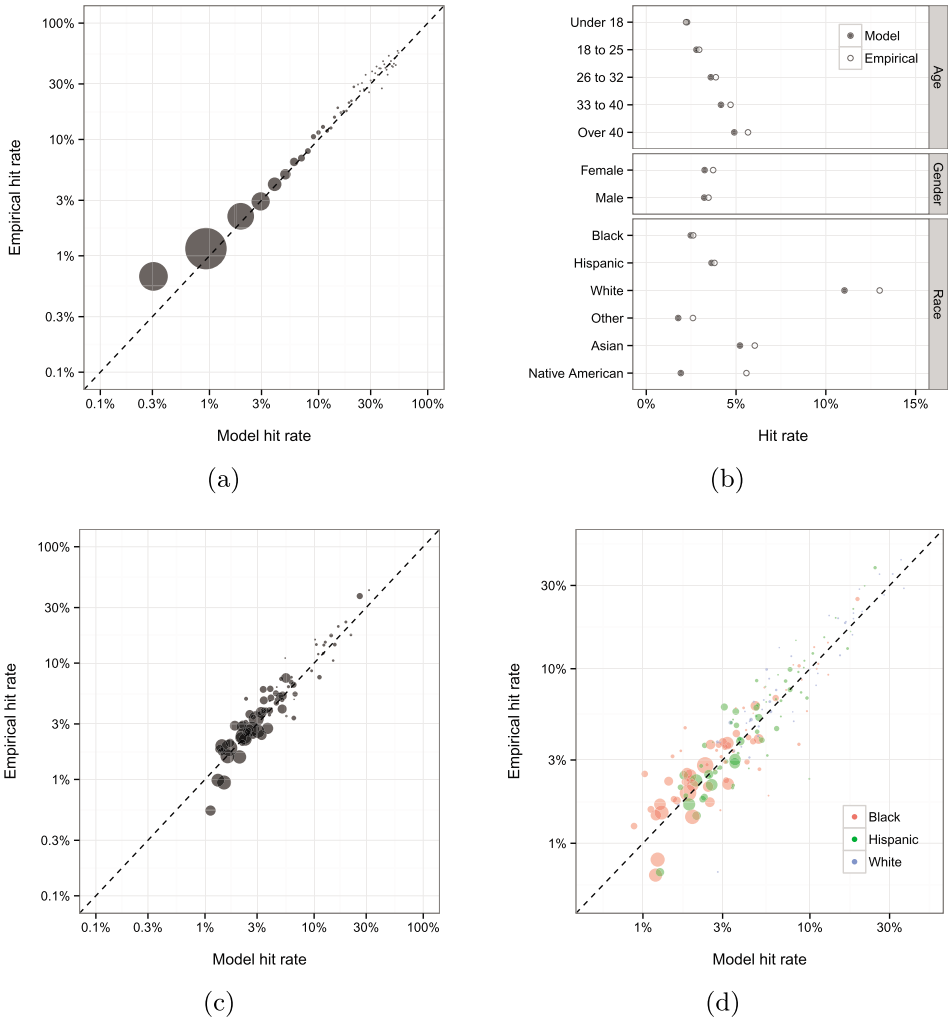

FIG. 8. *In Panel* (a), *stops are binned by model-predicted hit rate to the nearest percent. A point, sized by the number of stops, is plotted for each bin, comparing the model-predicted hit rate to the actual (empirical) hit rate. Panel* (b) *shows that for various values of the features age, race and gender, there is little difference between the model-predicted hit rate and the empirical hit rate. Panel* (c) *shows bin stops by precinct and again indicates that our model predicts well over the entire range of hit rates. In Panel* (d) *stops are binned by precinct and race, for stops of black, white and Hispanic suspects. For each bin with more than* 100 *stops, we plot a point, sized by the number of stops and shaded according to race, comparing the model-predicted hit rate to the empirical hit rate. In Panels* (a), (c) *and* (d) *the plotted points lie close to the dashed* 45 *degree line, indicating that our model predicts well over the entire range of hit rates and for interactions between important features.*

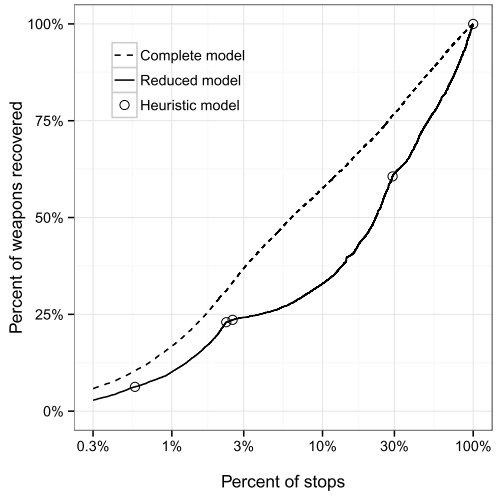

FIG. 9. *The estimated percentage of weapons recovered as a function of the number of stops conducted* (*note that the x-axis is on a log scale*) *for various models*.

APPENDIX B: AN ALTERNATIVE STOP HEURISTIC

In our primary analysis, we constructed stop heuristics that relied on area-specific thresholds. However, for social, political or legal reasons, one might prefer a policy that applies a uniform threshold across the city. To construct such a policy, we follow the procedure outlined in Section 3.5, but omit the precinct and location-type covariates in equation (3). Figure 9 plots the performance of this model. While there is again little difference between the reduced and heuristic models, both fare considerably worse than when location information is included.

APPENDIX C: GUN RECOVERY

Our main analysis considered stops with the suspected crime of criminal possession of a weapon (CPW), and examined whether or not a weapon was found. A weapon in this case may refer to a gun, knife or "other." We note that in 2008–2012, of the 27,000 stops where a weapon was discovered, a knife was found 77% of the time, whereas a gun was found only 12% of the time. Here we repeat the analyses in Sections 3.1 and 3.4 for gun recovery, training a model on CPW stops in 2009–2010 and estimating the ex ante probability that a gun will be found in each CPW stop between 2011–2012. In Figure 10(a) we see that 95% of CPW stops in 2011–2012 have less than a 1% ex ante chance of recovering a gun. Figure 10(b) shows the percentage of guns recovered if stops are conducted according to their model-estimated probability, from highest to lowest. The ma-

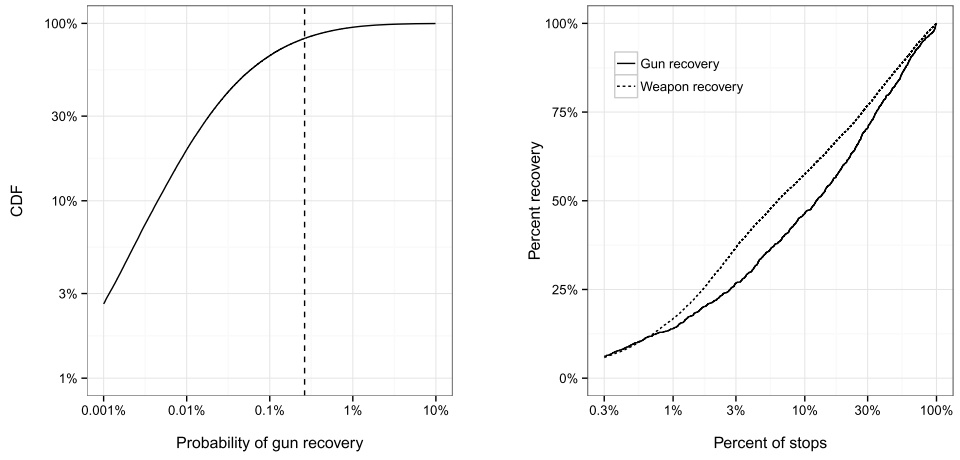

FIG. 10. *Panel* (a) *shows the distribution of the ex ante probability of finding a gun for CPW stops conducted in* 2011–2012 (*note that the x-axis is on a log scale*). 95% *of such stops have less than a* 1% *ex ante chance of turning up a gun, and the vertical line indicates the overall likelihood of finding a gun on a stopped suspect. Panel* (b) *plots the estimated percentage of guns recovered as a function of the number of stops conducted, where the stops are ordered by their model-predicted likelihood of turning up a gun, from highest to lowest. The best* 10% *of stops result in* 46% *of guns recovered, and the best* 50% *of stops result in* 83% *of guns recovered. The weapon recovery curve from Panel* 5(a) *is superimposed for comparison.*

jority of guns can be recovered by conducting only the 12% of stops with highest model-estimated probability of gun recovery.

APPENDIX D: RANDOM FOREST

We repeat the analyses in Sections 3.1 and 3.4 using a random forest classifier[21] trained on all CPW stops from 2009–2010. We estimated the ex ante probability of finding a weapon for CPW stops in 2011–2012, and find an AUC score of 0.83. Figure 11 demonstrates that the random forest classifier gives qualitatively similar results to the regression model used in the primary analysis.

[21]This classifier was implemented with Python's Scikit-learn package [Pedregosa et al. (2011)] using 1000 trees and a minimum of 10 samples required to split an internal node.

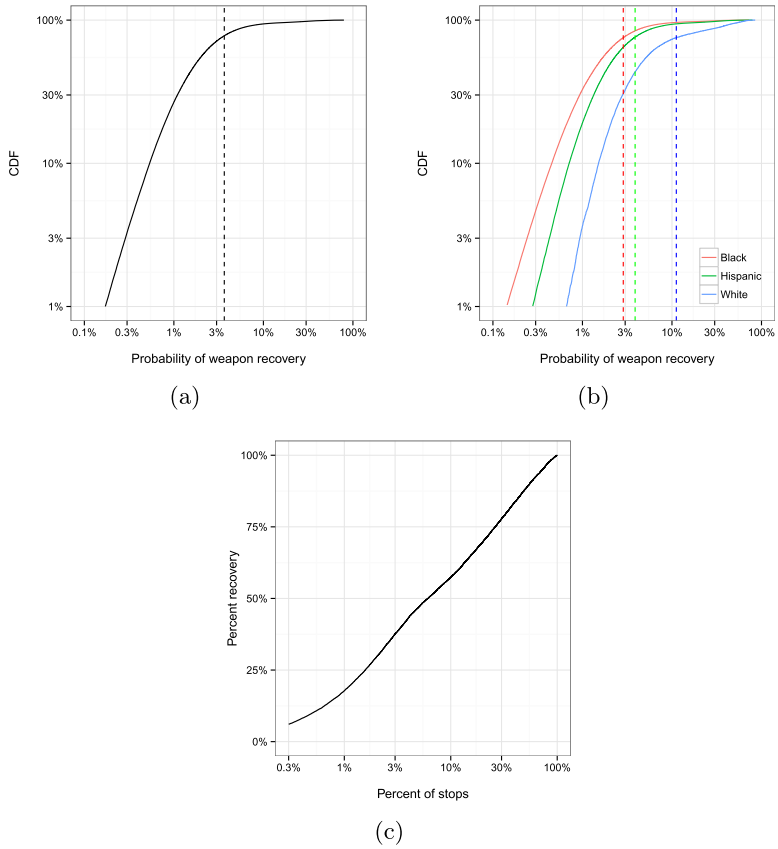

FIG. 11. *Panel* (a) *shows the distribution of the ex ante probability of finding a weapon for CPW stops conducted between* 2011–2012 (*note that the* x*-axis is on a log scale*), *calculated using a random forest model*, *and Panel* (b) *disaggregates this distribution by suspect race*. *Panel* (c) *plots the estimated percentage of weapons recovered as a function of the number of stops conducted*, *where the stops are ordered by their model-predicted likelihood of turning up a weapon*, *from highest to lowest*. *The best* 10% *of stops result in* 57% *of weapons recovered*, *and the best* 50% *of stops result in* 88% *of weapons recovered*.

REFERENCES

ANTONOVICS, K. and KNIGHT, B. G. (2009). A new look at racial profiling: Evidence from the Boston police department. *Rev. Econ. Stat*. **91** 163–177.

ANWAR, S. and FANG, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *Am. Econ. Rev.* **96** 127–151.

ARROW, K. (1973). The theory of discrimination. *Discrimination in Labor Markets* **3** 3–33.

AYRES, I. (2002). Outcome tests of racial disparities in police practices. *Justice Research and Policy* **4** 131–142.

BACH, F. R., HECKERMAN, D. and HORVITZ, E. (2006). Considering cost asymmetry in learning classifiers. *J. Mach. Learn. Res*. **7** 1713–1741. MR2274422

BECKER, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *J. Polit. Econ.* **101** 385–409.

BECKER, G. S. (2010). *The Economics of Discrimination*. Univ. Chicago press, Chicago, IL.

BERK, R. (2012). *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer Science & Business Media, Berlin.

BOTTOU, L. (1998). Online learning and stochastic approximations. *On-line Learning in Neural Networks* **17** 9–42.

COVIELLO, D. and PERSICO, N. (2013). An economic analysis of black–white disparities in NYPD's stop and frisk program. Available at http://www.nicolapersico.com.

CZERLINSKI, J., GIGERENZER, G. and GOLDSTEIN, D. G. (1999). How good are simple heuristics? In *Simple Heuristics That Make Us Smart* 97–118. Oxford Univ. Press, Oxford.

DANIELS ET AL. V. THE CITY OF NEW YORK (2001). 198 F.R.D. 409, 411, 422, S.D.N.Y.

DAVIS V. CITY OF NEW YORK (2013). No. 10 Civ. 0699.

DIGGLE, P. (1985). A kernel method for smoothing point process data. *Applied Statistics* **34** 138–147.

EBERHARDT, J. L., GOFF, P. A., PURDIE, V. J. and DAVIES, P. G. (2004). Seeing black: Race, crime, and visual processing. *J. Pers. Soc. Psychol.* **87** 876–893.

EPP, C. R., MAYNARD-MOODY, S. and HAIDER-MARKEL, D. P. (2014). *Pulled over: How Police Stops Define Race and Citizenship*. Univ. Chicago Press, Chicago, IL.

EWENS, M., TOMLIN, B. and WANG, L. C. (2014). Statistical discrimination or prejudice? A large sample field experiment. *Rev. Econ. Stat.* **96** 119–134.

FAGAN, J. and GELLER, A. (2014). Following the script: Narratives of suspicion in Terry stops in street policing. Columbia Public Law Research Paper 14-410.

FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S. and AMORIM, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133–3181. MR3277155

FLOYD V. CITY OF NEW YORK (2013). 959 F. Supp. 2d 540, S.D.N.Y.

GELMAN, A., FAGAN, J. and KISS, A. (2007). An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. *J. Amer. Statist. Assoc.* **102** 813–823. MR2411646

GIGERENZER, G. and GOLDSTEIN, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychol. Rev.* **103** 650.

GOEMAN, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biom. J.* **52** 70–84. MR2756594

GROGGER, J. and RIDGEWAY, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *J. Amer. Statist. Assoc.* **101** 878–887. MR2324089

ILLINOIS V. WARDLOW (2000). 528 U.S. 119.

KNOWLES, J., PERSICO, N. and TODD, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *J. Polit. Econ.* **109**.

KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411. MR2719657

LEGEWIE, J. (2016). Racial profiling in stop-and-frisk operations: How local events trigger periods of increased discrimination. *Am. J. Sociol.* To appear.

LERMAN, A. E. and WEAVER, V. (2014). Staying out of sight: Concentrated policing and local political action. *Ann. Am. Acad. Polit. Soc. Sci.* **651** 6–21.

LIGON V. CITY OF NEW YORK (2013). No. 12 Civ. 2274 (SAS).

LOVIE, A. D. and LOVIE, P. (1986). The flat maximum effect and linear scoring models for prediction. *J. Forecast.* **5** 159–168.

MICHIGAN DEPT. OF STATE POLICE V. SITZ (1990). 496 U.S. 444.

MILGRAM, A., HOLSINGER, A. M., VANNOSTRAND, M. and ALSDORF, M. W. (2015). Pretrial risk assessment: Improving public safety and fairness in pretrial decision making. *Federal Sentencing Reporter* **27** 216–221.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. and DUCHESNAY, E. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348

PERSICO, N. (2009). Racial profiling? Detecting bias using statistical evidence. *Annual Review of Economics* **1** 229–254.

RIDGEWAY, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *J. Quant. Criminol.* **22** 1–29.

RIDGEWAY, G. (2007). Analysis of racial disparities in the New York Police Department's stop, question, and frisk practices. Rand Corporation.

RIDGEWAY, G. and MACDONALD, J. M. (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *J. Amer. Statist. Assoc.* **104** 661–668. MR2751446

RUDOVSKY, D. and ROSENTHAL, L. (2013). Debate: The constitutionality of stop-and-frisk in New York City. *U. Pa. L. Rev. Online* **162** 117–117.

TERRY V. OHIO (1968). 392 U.S. 1.

USTUN, B. and RUDIN, C. (2014). Methods and models for interpretable linear classification. Preprint. Available at arXiv:1405.4047.

WILSON, J. Q. and KELLING, G. L. (1982). Broken windows. *Atlantic Monthly* **249** 29–38.

S. GOEL
MANAGEMENT SCIENCE AND ENGINEERING
STANFORD UNIVERSITY
475 VIA ORTEGA
STANFORD, CALIFORNIA 94305
USA
E-MAIL: scgoel@stanford.edu

J. M. RAO
MICROSOFT RESEARCH
641 AVENUE OF THE AMERICAS
7TH FLOOR
NEW YORK, NEW YORK 11249
USA
E-MAIL: justin.rao@microsoft.com

R. SHROFF
CENTER FOR URBAN SCIENCE AND PROGRESS
NEW YORK UNIVERSITY
1 METROTECH CENTER, 19TH FLOOR
BROOKLYN, NEW YORK 11201
USA
E-MAIL: ravi.shroff@nyu.edu